

Approaching allelic probabilities and Genome-Wide Association Studies from beta distributions.

J. Santiago García-Cremades^{*1}, Ángel del Río¹, José A. García², Javier Gayán^{3,4},
Antonio González-Pérez^{3,5}, Agustín Ruiz^{3,6}, Oscar Sotolongo-Grau⁶, and Manuel
Ruiz-Marín²

¹Departamento de Matemáticas, Universidad de Murcia.

²Departamento de Métodos Cuantitativos e Informáticos, Universidad Politécnica de Cartagena.

³Department of Estructural Genomics, Neocodex, Sevilla, Spain.

⁴Bioinfosol, Sevilla, Spain.

⁵Centro Andaluz de Estudios Bioinformáticos (CAEBi), Sevilla, Spain.

⁶Memory Clinic of Fundació ACE. Institut Català de Neurociències Aplicades. Barcelona. Spain.

^{*}Correspondence to: José Santiago García Cremades, Departamento de Matemáticas, Universidad de Murcia, 30100, Espinardo, Spain. E-mail: js.garciacremades@gmail.com

Abstract

In this paper we have proposed a model for the distribution of allelic probabilities for generating populations as reliably as possible. Our objective was to develop such a model which would allow simulating allelic probabilities with different observed truncation and degree of noise. In addition, we have also introduced here a complete new approach to analyze a genome-wide association study (GWAS) dataset, starting from a new test of association with a statistical distribution and two effect sizes of each genotype. The new methodological approach was applied to a real data set together with a Monte Carlo experiment which showed the power performance of our new method. Finally, we compared the new method based on beta distribution with the conventional method (based on Chi-Squared distribution) using the agreement Kappa index and a principal component analysis (PCA). Both the analyses show found differences existed between both the approaches while selecting the single nucleotide polymorphisms (SNPs) in association.

Keywords: GWAS; case-control study; allelic probability; beta distribution

1 Introduction

A fundamental point in a genome-wide association study (GWAS) is to define a model for allelic probabilities. A good model must strongly depend on the observed truncation and the degree of noise that distort the observed (empirical) distribution from the expected one. By controlling the truncation and the degree of noise, allelic probabilities can be simulated in a more reliable scenario.

The Human Genome Project (Lander et al. [2001]) and the successive improvements on physical and genetic maps of the human genome have boosted the post-genome era (Altshuler et al. [2010], Frazer et al. [2007], Sachidanandam et al. [2001]). Concomitant technological achievements in the genetic field have been successfully applied to uncover thousands of genetic variants linked to multiple phenotypes. The deep characterization of loci involved in both Mendelian and complex disorders will further help in improving diagnostic resolution and, ultimately, provide clues on the design of next generation therapeutics based on etiology, rather than in symptoms or clinical findings.

The genome-wide association studies (GWAS) appear to be unstoppable. The development of high density genome-wide panels of single nucleotide polymorphism (SNPs) and its application to bio-banked samples, accumulated during the last century, are the key elements explaining GWAS emergence. Till, date ongoing GWAS projects have published 1350 GWAS documents and more than 1800 GWAS significant loci (www.genome.gov/gwastudies; Freeze Dec/2012) (Hindorff et al. [2009]). The new loci have been detected using relatively well standardized methods based on linear additive models with or without covariants, a case-control design, by applying extensive quality control to raw genotyping data, by increasing density of markers based on inference of many non-genotyped markers using high performance computation (HPC), imputation techniques and by increasingly improved reference panels of single nucleotides polymorphisms (SNPs) (de Bakker et al. [2008]).

In spite of these successes, most GWAS findings, typically of small effect sizes, leave a large fraction of disease susceptibility still unexplained; a phenomenon commonly known as “the

case of the missing heritability” (Meesters et al. [2012]). Several potential explanations for this phenomenon were proposed (Manolio et al. [2009]). An excessive simplification of statistical methods applied to GWAS datasets might account for this problem. In this regard, allelic and additive models pervasively applied to GWAS data could be the genuine spherical cow (Shelton and Cliffe [2007]) on genetic research. Therefore, it is necessary to perform a continuous re-analysis and re-cycling of GWAS data by applying novel statistical methods to uncover those loci that match poorly with linear models (see Ruiz-Marín et al. [2010]).

Here, we have proposed a model for the distribution of allelic probabilities which allows to simulate allelic probabilities with different observed truncation and degree of noise. We have also introduced a complete new approach to GWAS analysis, starting from a new test for association and two effect sizes of each genotype. The new methodological approach was applied to a real data set together with a Monte Carlo experiment which showed the power performance of our new method and compared the new method with the conventional method.

2 Materials and Methods

2.1 Modeling Allelic Probability Distribution

Allele frequency refers to the proportion of a certain allele on a genetic locus of population. These proportions often exhibit extra variation that cannot be explained by a simple binomial distribution. The proportion (or binomial parameter p) does not remain constant in the course of collecting data. Considering the situation, it would be useful to assume that the binomial parameter p varies between observations. The data could be described assuming one of many continuous distributions for p , $0 < p < 1$. However, the most sensible distribution for p is the beta distribution, because it is the natural conjugate prior distribution in the Bayesian sense.

It is well known that most SNPs present a very low minor allele frequency (MAF), close to either 0 or 1 depending on the codification. However, these SNPs with very low MAF are systematically excluded from GWAS. After elimination, it is often assumed that the allele frequencies follow a uniform distribution. This is consistent with the beta approach, because the uniform distribution in the interval $[0,1]$ is the beta distribution with parameters $\alpha = \beta = 1$. However, the observed distribution of allelic frequencies is not quite uniform. Therefore, a finer analysis of the observed distribution is required.

The beta distribution with parameters α and β is denoted by $\text{Beta}(\alpha, \beta)$ and has the following probability density function (PDF)

$$f(a) = \begin{cases} \frac{B_{\alpha,\beta}(a)}{\int_0^1 B_{\alpha,\beta}(r)dr}, & \text{if } 0 < a < 1; \\ 0, & \text{otherwise.} \end{cases}$$

where $B_{\alpha,\beta}(a) = a^{\alpha-1}(1-a)^{\beta-1}$ and $0 < \alpha, \beta \leq 1$. The mean μ and variance σ^2 of $\text{Beta}(\alpha, \beta)$ are given by

$$\mu = \frac{\alpha}{\alpha + \beta}, \quad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \quad (1)$$

Notice that the parameters α and β are determined by the mean and variance (1) by the following formulae:

$$\alpha = \frac{\mu(\mu - \mu^2 - \sigma^2)}{\sigma^2}, \quad \beta = \frac{(1 - \mu)(\mu - \mu^2 - \sigma^2)}{\sigma^2}. \quad (2)$$

In order to define variables, we assumed the allelic probability as the probability that one of the possible alleles occurs. For this, we denoted the two most frequent alleles as A and B , respectively. This symbolic association was performed at random with equal probability. The allelic probability is defined as the probability of the occurrence of the allele denoted as A . This introduces a random variable, referred as *allelic probability* (\mathcal{AP}).

As explained above, since the values of \mathcal{AP} are proportions, the random variable \mathcal{AP} can be modeled with a beta distribution $\text{Beta}(\alpha, \beta)$. As the chosen allele is determined at random the mean of \mathcal{AP} should be 0.5. In terms of the beta distribution this implies that $\alpha = \beta$.

However, the beta distribution is not enough to properly model the allelic probability distribution in a real dataset. There are several considerations that we must take into account to explain this situation. For example, the commercial genome-wide SNP chips designs or quality controls (QC) applied to the genotyping studies, regularly exclude the SNPs with very small MAF, those with Hardy Weinberg disequilibrium, and those with a poor quality, etc. Furthermore, the incorporation of imputation methods may also introduce bias in genome-wide allelic distribution by imputing preferentially those SNPs with a higher MAF and those located in regions in strong linkage disequilibrium. Thus, one cannot assume that \mathcal{AP} takes all the values in the interval $[0, 1]$. Either design-based or QC-based pruning of SNPs induce a truncation of beta distribution.

Therefore, we may consider the random variable \mathcal{AP}_t of the truncated allelic probabilities in a GWAS with the truncation t , following a beta distribution truncated in some interval $[t, 1 - t]$, denoted as $\text{Beta}(\alpha, \alpha, t, 1 - t)$. Its probability density function is given by:

$$g(a) = \begin{cases} \frac{B_{\alpha, \alpha}(a)}{\int_t^{1-t} B_{\alpha, \alpha}(r) dr}, & \text{if } t < a < 1 - t; \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Moreover, universal or general truncation exists in commercial chips that have been designed based on data derived from worldwide populations as HapMap or 1000 genome project datasets. In contrast, the DNA samples for a specific study are obtained from local populations. It is well established that SNP frequencies vary geographically due to genetic drift and exceptionally, due to natural selection in specific endemic regions. Therefore, a low MAF in studied local population might not indicate necessarily an equivalent low MAF in worldwide populations. As a consequence, selected SNPs in a commercial chip might not be observed in the local population, yielding a very low MAF that displays allelic probability values close to 0 or 1. To further analyze this phenomenon, we introduced the random variable a priori truncated local allelic probability, \mathcal{LAP}_t , and the difference $\mathcal{D}_t = \mathcal{LAP}_t - \mathcal{AP}_t$, which measures the variation between the universal allelic probability and the local allelic probability of a given SNP.

2.1.1 The local allelic probabilities

We assumed that \mathcal{D}_t and \mathcal{AP}_t are independent random variables. To guarantee local allelic probabilities to be bound into the interval $[0, 1]$ (see Appendix A), we assumed that \mathcal{D}_t is truncated in the interval $[-t, t]$. Hence we modeled \mathcal{D}_t as a truncated normal distribution

$NT(0, \delta, -t, t)$. Its probability density function is given by:

$$h(x) = \begin{cases} \frac{n_{0,\delta}(x)}{\int_{-t}^t n_{0,\delta}(r)dr}, & \text{if } -t < x < t; \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

where $n_{0,\delta}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\delta^2}}$ is the probability density function of a Normal distribution with mean zero and standard deviation δ . Notice that the mean of \mathcal{D}_t is equal to zero and its variance is denoted by σ_D^2 .

As \mathcal{AP}_t and \mathcal{D}_t were assumed to be independent. The probability density function of

$$\mathcal{LAP}_t = \mathcal{AP}_t + \mathcal{D}_t \quad (5)$$

is the convolution of the probability density functions of \mathcal{AP}_t and \mathcal{D}_t .

As \mathcal{AP}_t is distributed as $\text{Beta}(\alpha, \alpha, t, 1-t)$, it is straightforward that its mean should be 0.5. The variance of \mathcal{AP}_t was calculated using the incomplete beta function and the regularized incomplete beta function as shown in Appendix B (19). On the other hand, as \mathcal{D}_t follows a $NT(0, \delta, -t, t)$ distribution, its mean value is zero and its variance is given by (6).

$$\sigma_D^2 = \delta^2 \left(1 - \frac{2\frac{t}{\delta}\phi(\frac{t}{\delta})}{2\Phi(\frac{t}{\delta}) - 1} \right), \quad (6)$$

where ϕ and Φ are the PDF and CDF of the standard normal distribution $N(0, 1)$ respectively. Moreover, we have $0 \leq \sigma_D^2 \leq \frac{t^2}{3}$ (see Lemma B.1).

As \mathcal{AP}_t and \mathcal{D}_t are independent, the variance of $\mathcal{LAP} = \mathcal{AP}_t + \mathcal{D}_t$ is $\sigma_l^2 = \sigma_u^2 + \sigma_D^2$ and hence

$$\sigma_u^2 \leq \sigma_l^2 \leq \sigma_u^2 + \frac{t^2}{3}.$$

This last expression includes the variance, taking into the account the noisy data. Indeed, if a higher degree of noise exists, the \mathcal{D}_t variance can be taken as its maximum value, $\sigma_D^2 = \frac{t^2}{3}$. So, in this case, the expression $\sigma_l^2 \approx \sigma_u^2 + \frac{t^2}{3}$ can be used as a good approximation for \mathcal{LAP} variance.

2.1.2 A real data example

In order to illustrate our approach we took the data from a practical case. The GWAS dataset is an imputed GWAS with 1,237,567 SNPs and 1225 individuals genotyped by the Translational Genomics Research Institute (TGEN) (Reiman et al. [2007]), previously processed as part of a genome-wide meta-analysis looking for Alzheimer's disease genetic risk factors (Antúnez et al. [2011, May 31]).

Figure 1 shows the comparison between the empirical allele frequencies with the PDF of the uniform (A), beta (B) and truncated beta (C) distributions, where the empirical allele frequencies are represented by dots and the PDFs are represented by continuous lines. The parameters for the beta and truncated beta distribution were taken for the distribution to have the same mean and variance as the empirical data. Figure 1 shows that the beta distribution is not good enough to model the empirical allelic frequencies. The observed divergence mainly

occurs in the two tails of the distribution where the theoretical beta distribution increases, while the data dramatically decreases. The observed truncation is attributed to the removal of low MAF alleles during QC and the relative inability of imputation methods to make good inferences for low MAF SNPs.

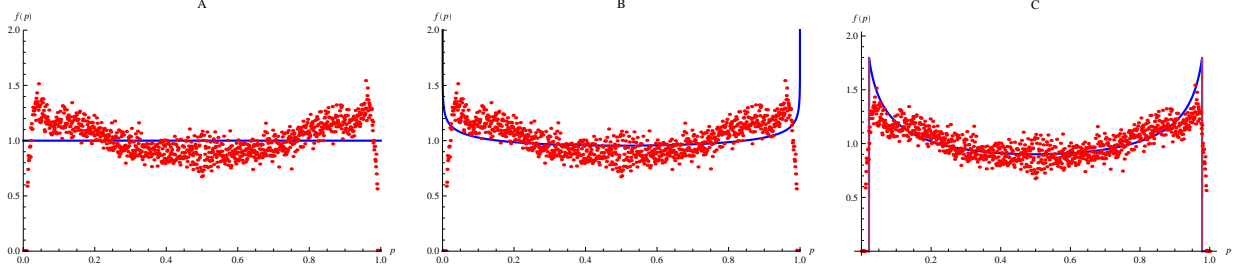


Figure 1: Empirical relative frequencies (dots) and uniform (A), beta (B), and truncated beta (C) PDFs (continuous lines). (A) PDF of uniform distribution in $[0, 1]$. (B) PDF of Beta (α, β) . (C) PDF of Beta $(\alpha, \beta, t, 1 - t)$. The parameters α, β of the beta distributions are taken to fit the mean and variance of empirical relative frequencies ($\alpha = \beta = 0.928$) and the truncation is computed as the percentile 1 ($t = 0.0218$). The empirical frequencies have been calculated grouping the data in 1000 intervals.

In order to fit the data to equation (5), we took t as the percentile 1 of the allelic probabilities: $t = 0.0218$. Since σ_D^2 belongs to the interval $[0, \frac{t^2}{3}]$ so that we can estimate σ_D^2 as if it is a uniform random variable in such an interval. This means that the expected value of σ_D^2 is calculated as $\sigma_D^2 = \frac{t^2}{6} = 7.9 \cdot 10^{-5}$. Then, the parameter δ can be estimated as $\delta = 9.59 \cdot 10^{-3}$ from (6).

Since t is known, σ_l can be estimated from the sample. As σ_u only depends on the parameters α and t , the value of α can be computed by solving (see Appendix B),

$$\sigma_u^2 = \sigma_{\alpha,t}^2 = \frac{1}{(4\alpha + 2)B(\alpha, \alpha)} \frac{t^\alpha(1-t)^\alpha(4t-2)}{1-2I_t(\alpha, \alpha)} + \frac{\alpha+1}{4\alpha+2} - \frac{1}{4} = \sigma_l^2 - \frac{t^2}{6}, \quad (7)$$

where $\sigma_{\alpha,t}^2$ is the variance of the truncated beta distribution. Therefore, $\alpha = 0.7199$ is obtained from this equation. The three parameters, t , δ , and α determine the \mathcal{LAP}_t distribution fitting to the dataset. Figure 2 shows the empirical relative frequencies (dots) and the model of allelic probabilities \mathcal{LAP}_t (continuous line) with the parameters computed above.

The \mathcal{LAP}_t model fits much better to a real dataset than the previous analyzed models. This is especially true for the tails of the probability distribution functions. Besides, the mean squared error (MSE) estimator decreases for the new model. As shown in Table I, the minimum MSE is reached for the \mathcal{LAP}_t model with 70.16% decrease in the MSE when compared with the uniform model.

Table I: The Mean Squared Error calculated in the four different distributions described fitting the empirical allelic probability grouped in 1000 intervals.

TGEN AP	Uniform	Beta	Beta _t	\mathcal{LAP}_t
<i>MSE</i>	0.0429	0.0517	0.0315	0.0128

2.2 Case-control probabilities

We assumed a GWAS, where a SNP was typed for N individuals, with N_0 controls and N_1 cases. M_s denotes the number of individuals which present a certain genotype (s).

Next we attempted to test for the following null hypothesis,

$$H_0 : \text{the absence of association in genotype } s. \quad (8)$$

Notice that under this null, the probability of being case conditioned to having a given genotype s , b_s , is the same as the probability of being case, b . In order to test this null, the following binomial distribution, $\mathbf{Bin}(M_s, b_s)$, consisting of number of cases with genotype s can be considered. Under the null, b_s can be estimated as $\hat{b}_s = \hat{b} = N_1/N$. On the other hand, as we mentioned before that the most sensible distribution for b_s is the beta distribution, $\text{Beta}(\alpha_{M_s}, \beta_{M_s})$. This is based on the fact that the beta distribution is the conjugate prior of the binomial distribution (MacKay [2003]).

The mean of b_s can be estimated by $\mu = \hat{b}_s = \frac{N_1}{N}$ and its variance $\sigma^2 = \frac{\hat{b}_s(1-\hat{b}_s)}{M_s} = \frac{N_0 N_1}{N^2 M_s}$. Since the population under study is finite it is necessary to adjust the variance σ^2 for the population size with the finite population correction factor $\frac{N-M_s}{N-1}$ (Isserlis [1918]). This is specially required when sample size M_s is not small in comparison with the population size N , so that $M_s > 0.05N$. Therefore, when $M_s > 5\%N$, the variance σ^2 remains as

$$\sigma^2 = \frac{N_0 N_1 (N - M_s)}{N^2 M_s (N - 1)}.$$

Thus, under the null hypothesis, we can estimate α_{M_s} and β_{M_s} ,

$$\alpha_{M_s} = \frac{N_1}{N} \left(\frac{M_s(N-1)}{N-M_s} - 1 \right) \quad \text{and} \quad \beta_{M_s} = \frac{N_0}{N} \left(\frac{M_s(N-1)}{N-M_s} - 1 \right). \quad (9)$$

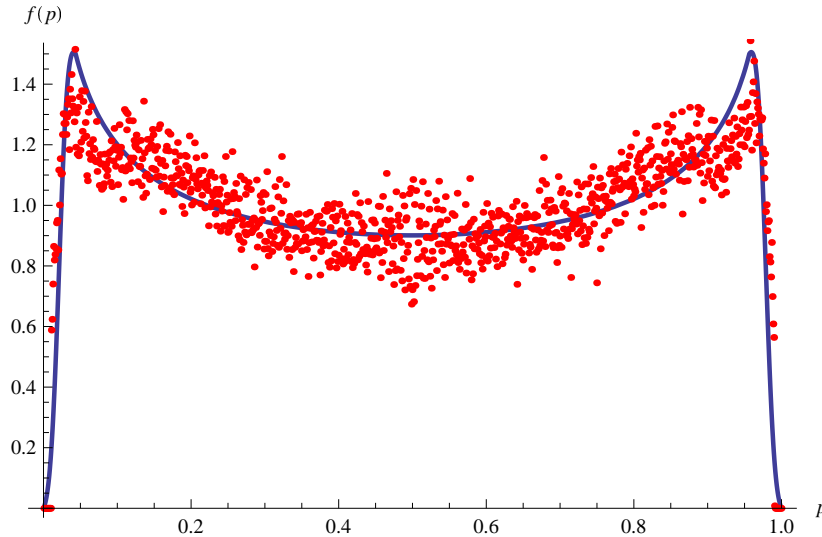


Figure 2: Empirical relative frequencies (dots) of allelic probabilities of the TGEN sample, with 1,237,567 SNPs and PDF of truncated $\text{Beta}(\alpha, \alpha, t) + \text{NT}(0, \delta, t)$ (continuous line) with 1225 individuals, where 757 are cases and 468 controls. The empirical relative frequencies were estimated grouping the data in 1000 intervals between $[0, 1]$.

Figure 3 compares the empirical relative frequencies of b_s in TGEN dataset (dots) with the PDF of $\text{Beta}(\alpha_{M_s}, \beta_{M_s})$ (continuous line) i.e., the beta distribution of b_s under the null hypothesis, for several values of M_s .

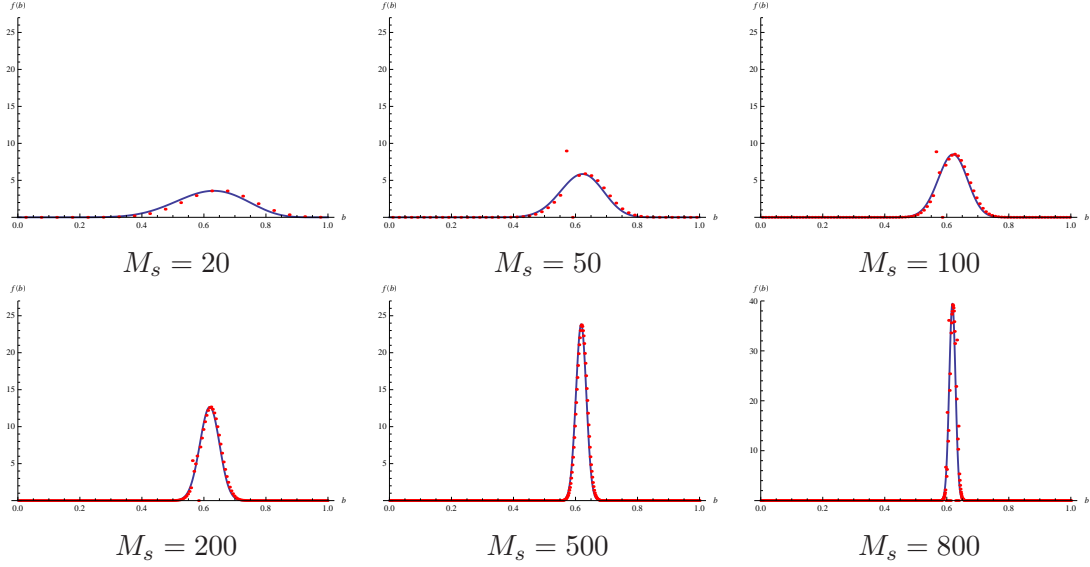


Figure 3: Empirical relative frequencies of b_s (dots) for genotypes with different numbers of individuals (20, 50, 100, 200, 500 and 800 of individuals) and PDF of $\text{Beta}(\alpha_{M_s}, \beta_{M_s})$ (continuous line).

Thus, we obtained a decision rule for H_0 at a desired confidence interval (CI) as,

$$\begin{aligned} &\text{Accept } H_0 && \text{if } q_{\epsilon/2} \leq b_s \leq q_{1-\epsilon/2} \\ &\text{Reject } H_0 && \text{otherwise} \end{aligned} \tag{10}$$

where ϵ is the type I error, $q_{\epsilon/2}$ and $q_{1-\epsilon/2}$ the extreme p-values, and both are tied by the beta distribution as,

$$\epsilon/2 = \Pr(\text{Beta}(\alpha_{M_s}, \beta_{M_s}) < q_{\epsilon/2}) = \Pr(\text{Beta}(\alpha_{M_s}, \beta_{M_s}) > q_{1-\epsilon/2}). \tag{11}$$

Let p denote the proportion of individuals of the general population which presents a given phenotype, and let c_s denote the same proportion calculated in the individuals of a sample with a given genotype s . The ratio $\varphi_s = c_s/p$ is called the effect of the genotype s on the phenotype in a given sample and represents the proportion in which an individual with genotype s has more probability of presenting a phenotype than general population. However, controls are a representation of the population, while cases are a sample of individuals with the phenotype. Therefore, it is possible to use the GWAS to find and estimation for c_s , (see Appendix C). Also, φ_s and b_s can be related as

$$\varphi_s = \frac{N_0}{N_1} \frac{b_s}{1 - b_s}, \quad b_s = \frac{\varphi_s N_1}{N_0 + \varphi_s N_1}. \tag{12}$$

These last expressions allow constructing a decision rule that takes into account the effect (φ_s) of the genotype s .

The new decision rule is written in the same way as (10) but in this case,

$$\epsilon/2 = Pr(\text{Beta}(\xi_0, \xi_1) < q'_{\epsilon/2}) = Pr(\text{Beta}(\xi_0, \xi_1) > q'_{1-\epsilon/2}), \quad (13)$$

where,

$$\xi_0 = \frac{\varphi_s N_1}{N_0 + \varphi_s N_1} \left(\frac{M_s(N-1)}{N-M_s} - 1 \right) \text{ and } \xi_1 = \frac{N_0}{N_0 + \varphi_s N_1} \left(\frac{M_s(N-1)}{N-M_s} - 1 \right). \quad (14)$$

The measured effect, applying the decision rule (10) (φ_s), takes only into account the data contained in the sample. However, the effect (φ_s) that delimits the rejection region for a given confidence level ϵ can be computed from decision rule (13). Here, we called this effect as the *critical effect* of genotype s , with a certain confidence level ϵ and it is denoted as Ψ_s^ϵ (Appendix C).

2.2.1 Exploring BT in different genetic models

The test for GWAS explained above, hereafter, would be referred as the Beta Test (BT) that considers each genotype characteristic independently. This approach provides some new possibilities with respect to the conventional association test based on contingency tables and hereafter, would be called as Conventional Test (CT). In univariate CT, one can consider five genotype characteristics corresponding to the two alleles (A and B) and their three possible combinations (AA , AB and BB), separately.

Many contingency tables can be constructed in the CT by considering two or three genotype characteristics. Similarly, many models can be used for the BT, considering the different non-empty subsets of the genotype characteristic. For example, there are eight possibilities for both univariate CT and BT. However, most of the models are not meaningful for genetic purposes. For example, the allelic feature against the homozygous genotype or the heterozygous presence compared to the dominant model are two comparisons that do not provide useful information. On the other hand, the odds-ratio is not well defined when more than two characteristics are considered.

Therefore, in practice, only five models are used, namely *allelic* (A versus B), *dominant* ($AA \cup AB$ versus BB), *recessive* (AA versus $AB \cup BB$), *heterozygous* ($AA \cup BB$ versus AB) and *genotype* (AA versus AB versus BB).

Next, we focused our attention in the allelic CT model and the alleles A and B of BT models because of the fact that the allelic CT model is the most common model used in GWAS. The relationships among the models can be seen in Appendix D. The remaining models are available from the authors upon request.

2.3 Measuring (or relating) locus effect in BT models

Conventionally, odds-ratio is the effect measure estimated in a case-control GWAS. Let OR_a be the odds-ratio in allelic model between cases and controls. The magnitude of effect of the allelic CT model (OR_a) can also be related to the effect of the BT. In other words, the effect of a given model can be expressed as a function of OR_a .

In order to simulate a case-control population, it is necessary to fix an odds-ratio value, for example, OR_a , and another parameter like the MAF of a population of controls. Let x_0 be the

MAF in a population of controls, and x_1 the MAF in a population of cases. If OR_a is fixed, x_1 can be written as

$$x_1 = \frac{OR_a \cdot x_0}{1 - x_0 \cdot (1 - OR_a)}. \quad (15)$$

Next, using the Hardy-Weinberg equilibrium (for short, HWE), it is possible to simulate the case-control population of SNPs.

The odds-ratio, as usually understood, shows the difference between the probability distribution for being case and the probability distribution for being control. The BT assumed a different null hypothesis: in each genotype characteristic, the probability of being case is contained in a confidence interval centered around the proportion between cases and total population $N_1/(N_0 + N_1)$. This proportion can be measured with the effect φ_s , and it is fixed for each genotype characteristic. From the HWE and using (15), the OR_a and φ_s (for $s = A$ or B) are related as follows:

$$\varphi_A = \frac{OR_a}{(1 - x_0 \cdot (1 - OR_a))} \quad \text{and} \quad \varphi_B = \frac{1}{(1 - x_0 \cdot (1 - OR_a))}. \quad (16)$$

Remarkably, if allelic frequencies in cases and controls are equal, there is no effect in any model. In other words, $x_0 = x_1$ implies that $OR_a = 1$ and $\varphi_s = 1$. Figure 4 shows the effects (φ_s) for different models as a function of the allelic odds-ratio ($OR_a \in [0, 10]$).

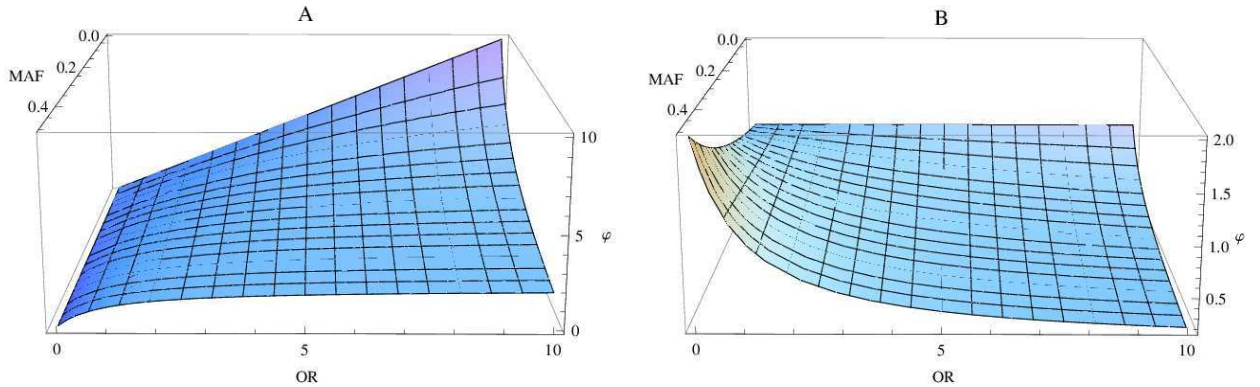


Figure 4: In the 3D graphic A, there are the effect of $s = A$ (φ_A) respect of the OR_a between zero and 10. In the 3D graphic B, there are the effect of $s = B$ (φ_B) respect of the OR_a between zero and 10, for different values of MAF .

3 Monte Carlo Simulation

A simulated dataset, consisting of $N_1 = 1000$ cases and $N_0 = 1000$ controls, was analyzed with both conventional chi-square (CT) and the new association (BT) tests in allelic models. A total of 10^5 SNPs, with different minor allele frequencies ($0 \leq MAF \leq 0.5$), were simulated under the null hypothesis i.e., to have no effect on the trait ($OR_a = 1$ or all the effects are equal to 1). This analysis can reflect whether the new Beta Test (BT) conforms to the theoretical beta distribution, and compares how the Conventional Test (CT) conforms to the Chi-square distribution. For estimating the type I error for each test (size of the test), we counted the number of test-statistics that had values above the critical values of the expected distribution.

Table II: Type I Error (size) for each model of Beta and Conventional tests, respectively, taking the MAF as a random value between $[0, 0.5]$ and different critical values of the expected distribution with corresponding confidence levels (0.05, 0.01 and 0.001).¹

Expected	BT							
	Allele A	Allele B	DomA	DomB	Homoz	s=AA	s=AB	s=BB
0.05	0.0489	0.0491	0.0493	0.0476	0.0491	0.0496	0.049	0.0494
0.01	0.00989	0.00989	0.00986	0.00926	0.0103	0.00996	0.0104	0.00982
0.001	0.000978	0.000957	0.00101	0.000739	0.000926	0.00107	0.000968	0.000989

Expected	CT				
	ALLELIC	DOM	REC	HETEROZ	GENO
0.05	0.0491	0.0494	0.0477	0.0492	0.0466
0.01	0.00989	0.00981	0.00926	0.0103	0.00937
0.001	0.000957	0.000978	0.000739	0.000916	0.000947

¹ The models are composed in BT by Allele A, Allele B, Dominant of A ($AA \cup AB$), Dominant of B ($AB \cup BB$), Homozygous ($AA \cup BB$) $s = AA$, $s = AB$ and $s = BB$. While in CT, the models are *allelic* (A versus B), *dominant* ($AA \cup AB$ versus BB), *recessive* (AA versus $AB \cup BB$), *heterozygous* ($AA \cup BB$ versus AB) and *genotype* (AA versus AB versus BB).

In order to simulate the case-control populations under certain conditions, only two parameters (MAF and an odds-ratio or φ_s) need to be fixed, as explained before. Indeed, fixing MAF in controls and the effect (odds-ratio or φ_s), the entire population is already defined.

The populations were simulated by generating at each SNP a random value for genotypic probabilities of controls according to MAF value and HWE. Once controls' MAF value and effect were fixed, the MAF values for cases were straightforwardly obtained, and following HWE, the genotype probabilities were computed, generating the cases population.

Using this method, both tests, CT and BT, yielded approximately the expected number of false positives (see Table II), suggesting they all conform to the expected theoretical distributions.

To estimate the power of both CT and BT, we carried out an analysis of a simulated dataset of 1000 cases and 1000 controls. Sets of 10^5 SNPs were simulated under different alternative hypothesis (see Table III), with different effect sizes and minor allele frequencies (0.05, 0.2 and 0.4), with a fixed confidence level $\epsilon = 0.05$. The effect of the genotype characteristics is related with the odds-ratio of corresponding model. Note that some scenarios could not be estimated due to HWE restrictions. Our results showed that both tests detected SNPs in association with similar power, depending on the new effect measurement and presenting there the allelic models comparison.

Table III: Power of both tests BT and CT in simulated data, taking SNPs under the alternative of association, with different minor allele frequencies (MAF) and effect sizes at allelic models (φ_A and φ_B).¹

Allele A		MAF = 0.05 $\varphi_A \in [0, 20]$		MAF = 0.2 $\varphi_A \in [0, 5]$		MAF = 0.4 $\varphi_A \in [0, 2.5]$	
Fixed φ_A	MODEL	POWER (in %)	OR_a	POWER (in %)	OR_a	POWER (in %)	OR_a
0.5	BT A	98.8	0.49	100	0.44	100	0.38
	CT ALLELIC	98.89		100		100	
0.67	BT A	74.4	0.66	99.99	0.62	100	0.55
	CT ALLELIC	75.65		99.99		100	
0.8	BT A	31.79	0.79	91.12	0.76	99.96	0.71
	CT ALLELIC	33.27		91.12		99.96	
1.25	BT A	40.26	1.27	96.7	1.33	100	1.5
	CT ALLELIC	40.21		96.7		100	
1.5	BT A	90.82	1.54	100	1.71	100	2.25
	CT ALLELIC	90.75		100		100	
2	BT A	100	2.11	100	2.67	100	6
	CT ALLELIC	100		100		100	

Allele B		MAF = 0.05 $\varphi_B \in [0, 1.05]$		MAF = 0.2 $\varphi_B \in [0, 1.25]$		MAF = 0.4 $\varphi_B \in [0, 1.67]$	
Fixed φ_B	MODEL	POWER (in %)	OR_a	POWER (in %)	OR_a	POWER (in %)	OR_a
0.8	BT B	100	6	100	2.25	100	1.63
	CT ALLELIC	100		100		100	
0.9	BT B	100	3.22	100	1.56	96.97	1.28
	CT ALLELIC	100		100		96.97	
0.95	BT B	100	2.05	86.36	1.26	48.67	1.13
	CT ALLELIC	100		86.36		48.7	
1.05	BT B	100	0.048	90.96	0.76	49.51	0.88
	CT ALLELIC	100		90.96		49.53	
1.1	BT B	NA	doesn't exist	100	0.55	97.62	0.77
	CT ALLELIC	NA		100		97.63	
1.2	BT B	NA	doesn't exist	100	0.17	100	0.58
	CT ALLELIC	NA		100		100	

¹ First table is for allele A model and the second table for allele B model. For each effect size φ and each minor allele frequency MAF the odds ratio of the allelic model OR_a is reported using (16), at a fixed confidence level $\epsilon = 0.05$.

4 Application to Real Data

In order to apply the BT approach to real data, a GWAS dataset with 1,237,567 SNPs (where $N_0 = 468$ controls and $N_1 = 757$ cases) was used (Section 2.1.2).

Table IV describes the parameters of the allelic probability distribution for cases, controls, and total population. As already shown in Figure 2, there are no allele frequencies lower than 0.01; however, there exists some noise close to the extreme values.

The calculated probability of being case with a genotype characteristic s is very sensitive to parameter M_s . This is shown in Figure 3, where b_s is plotted for different values of M_s . This implies that results depend on the frequency of having the genotype s and the sample size. While making inference on the correlation of a phenotype and a genotype, one should take into

Table IV: Summary of the TGEN allelic probabilities sample for cases, controls, and cases and controls.

TGEN_impQC2: 1,237,567 SNPs, 1225 individuals, 757 cases and 468 controls.			
	Cases & Controls	Cases	Controls
Empirical mean $\hat{\mu}$	0.500065	0.500063	0.500065
Empirical variance \hat{v}	0.0916691	0.0917528	0.0917662
Minimum \hat{v}	0.01	0.01	0.01
Truncation P_1	0.018605	0.018092	0.018233
$\hat{\alpha}$	0.666681	0.671064	0.669102
$\hat{\delta}$	0.01871	0.01561	0.01623
Mean Squared Error	0.0128	0.0154	0.013

account the number of individuals with such a genotype in the GWAS.

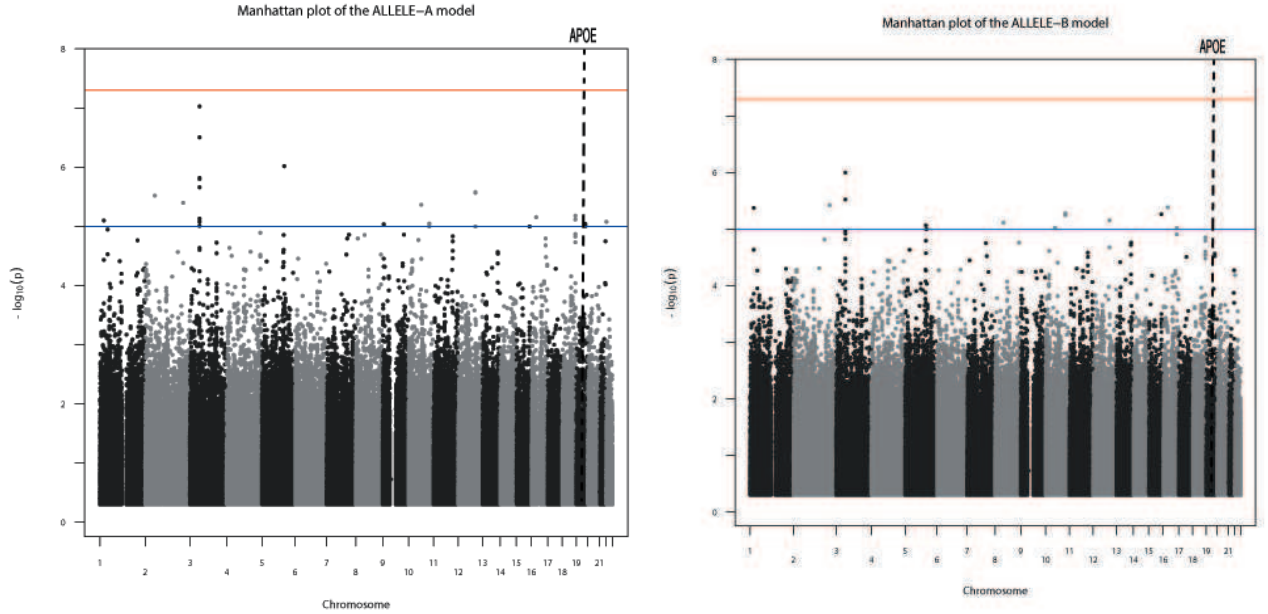


Figure 5: Manhattan plot of the new GWAS proposed with both alleles model in BT.

Figure 5 shows the Manhattan plot applied to the BT in the Allele A model within the sample described above (TGEN). Associated SNP on chromosome 19 is not displayed in the figure because of the chosen scale. This SNP is related to the apolipoprotein E (APOE) with a significant p -value = $1.33 \cdot 10^{-42}$ on BT.

Note that the CT is a one-tailed test where the null hypothesis of no association is rejected if p -value is lower than ϵ and the direction of the effect determines the risk or protective role of the SNP. However the BT is a two-tailed test, where the null hypothesis of no association is rejected if p -value is lower than $\epsilon/2$ or bigger than $1 - \epsilon/2$. In this case, the risk or protective role of the SNP was determined not only by the direction of the effect but also by the region of rejection. In other words, $p\text{-value} \leq \epsilon/2$ implies $\varphi_s \geq 1$ and *vice-versa*. Therefore, the p -value

information is enough to define the risk or protection association.

4.1 Measuring the concordance between CT and BT

The agreement between two sets of results could be measured by the Kappa index (K) of agreement (Cohen [1960]). In this case, we used two categories: being in association or not, with different levels of confidence. Kappa index is the estimator of agreement, compared in the paired models. At perfect agreement, K equals to one, while agreement given by chance gives a value of K close to zero.

Table V presents the allelic model comparison with their corresponding BT models (Allele A and Allele B). Notice that BT null hypothesis rejecting region contains approximately the expected number of significant SNPs than its corresponding confidence level. Nonetheless, there are some SNPs that reach significance in BT, but not in CT.

Table V: The allelic paired models compared with the Kappa index of agreement, calculated in column $Kappa$, and the number of SNPs in concordance where both tests detect or not an association with the phenotype with a confidence level ϵ . ¹

K1		ALLELE A				
	cl	Kappa	Pos Conc	CT-BT+	CT+BT-	Neg Conc
ALLELIC	$\epsilon = 0,05$	0.97	61861	1944	1651	1172111
	$\epsilon = 0,01$	0.94	11792	914	576	1224285
	$\epsilon = 0,001$	0.777	946	314	228	1236079
K2		ALLELE B				
	cl	Kappa	Pos Conc	CT-BT+	CT+BT-	Neg Conc
ALLELIC	$\epsilon = 0,05$	0.993	63164	456	348	1173599
	$\epsilon = 0,01$	0.985	12263	266	105	1224933
	$\epsilon = 0,001$	0.976	1163	46	11	1236347

¹ *Pos Conc* is the positive concordance, where both BT and CT reject the null hypothesis of no association. Similarly, *Neg Conc* is the negative concordance, where both BT and CT accept the null hypothesis. *BT+CT-* (respectively *BT-CT+*) represents the number of SNPs in discordance on both tests, where BT refuse (respectively, accept) an association and CT accept (respectively, reject) it.

When comparing CT with BT model, the Kappa index showed some differences in the association tests. For instance, a remarkable concordance ($K = 0.976$, $\epsilon = 0.001$) was observed while comparing the CT Allelic model with the BT Allele-B model (K2, Table V).

A remarkable result from Table V summarized in Table VI, reveals that all the SNPs, detected in association with the phenotype in CT test, were also detected in association with the BT test in any allele models (*A* or *B*). However, several SNPs detected being associated with the BT (in Allele *A* or *B* model) were not detected by the CT (see Table VI). For example, for a given confidence level $\epsilon = 0.001$, the 23% of positives, that are SNPs refusing H_0 , were detected by BT allelic models since CT allelic model could not detect them.

Table VI: Summary of the number of SNPs detected with association in CT Allelic model and both BT allele models (A and B) and their differences, with a confidence level ϵ .¹

Either allele A or B			
	Pos Conc	CT- BT+	CT+ BT-
$\epsilon = 0.05$	63512	2273	0
$\epsilon = 0.01$	12368	1137	0
$\epsilon = 0.001$	1174	354	0

¹ *Pos Conc* is the positive concordance, where both BT and CT reject the null hypothesis of no association. *BT+CT-* (respectively *BT-CT+*) represents the number of SNPs in discordance on both tests, where BT refuse (respectively, accept) an association and CT accept (respectively, reject) it.

The difference between BT results and the corresponding CT can be surprising at times. Indeed, one would expect much more concordance among CT and BT estimates. In order to understand the observed differences, a comparison of parameters space is advisable.

4.2 Comparisons with Principal Component Analyses

Principal Components Analysis (PCA) is a measurement that shows the relationship between two sets of parameters. We performed here PCA for analyzing the relationship between the parameters from CT (p -value of the χ^2 distribution and the odds-ratio OR_i corresponding to the model) and from BT (p -value of the beta distribution, the effect φ_s and the critical effect Ψ_s^ϵ). PCA finds which parameters explain the maximum variability and also sorts the components (transformed variables) by their explained variance; the original variables have corresponding weights in each components (Pearson [1901]).

In summary, PCA found which variables explain the maximum variability in the CT and BT results, the percentage of explained variance and the intensity.

Different variability explanation between allelic CT model and both alleles BT models using three principal components is shown in Figure 6 (a. and b.). These figures show the proportion explained variance is contributed by each parameter in the principal components. The p -value and φ_s of the BT, and the odds-ratio OR of the CT explain the same variance, which means these three parameters have the same direction as PC1. As expected, the p -value of CT (called as *pCHI*) explains the variance along the second principal component. However, the *Critical effect* of the BT Ψ_s^ϵ complements the third principal component, which means that some variability cannot be explained without it. As a matter of fact, the Critical effect contributes to the 87.72% of the third principal component, which explains the 15.17% of the total variance.

Therefore, the principal component analysis reflects that the results of both methods must differ in a region of the explaining variance.

4.3 Ranking of association

Top results for BT analysis are presented in Table VII. Briefly, we displayed the ranking of the smallest p -values in BT ($< 10^{-5}$) including either A or B allele estimation and their corresponding results using conventional one degree of freedom chi-squared tests applied to MAF (CT). We also included the ranking order observed for each marker using both approaches (BT and CT). As expected, most SNPs, but not all of them, exhibited very similar p -values and ranking order. This result is fully compatible with the performance of the global kappa index (Table V).

A whole and comprehensive ranking is also included (Supplementary file “RankingTGEN-pvalueBT.csv”). Of course, both strategies identified SNP marker rs4420638 on chromosome 19,

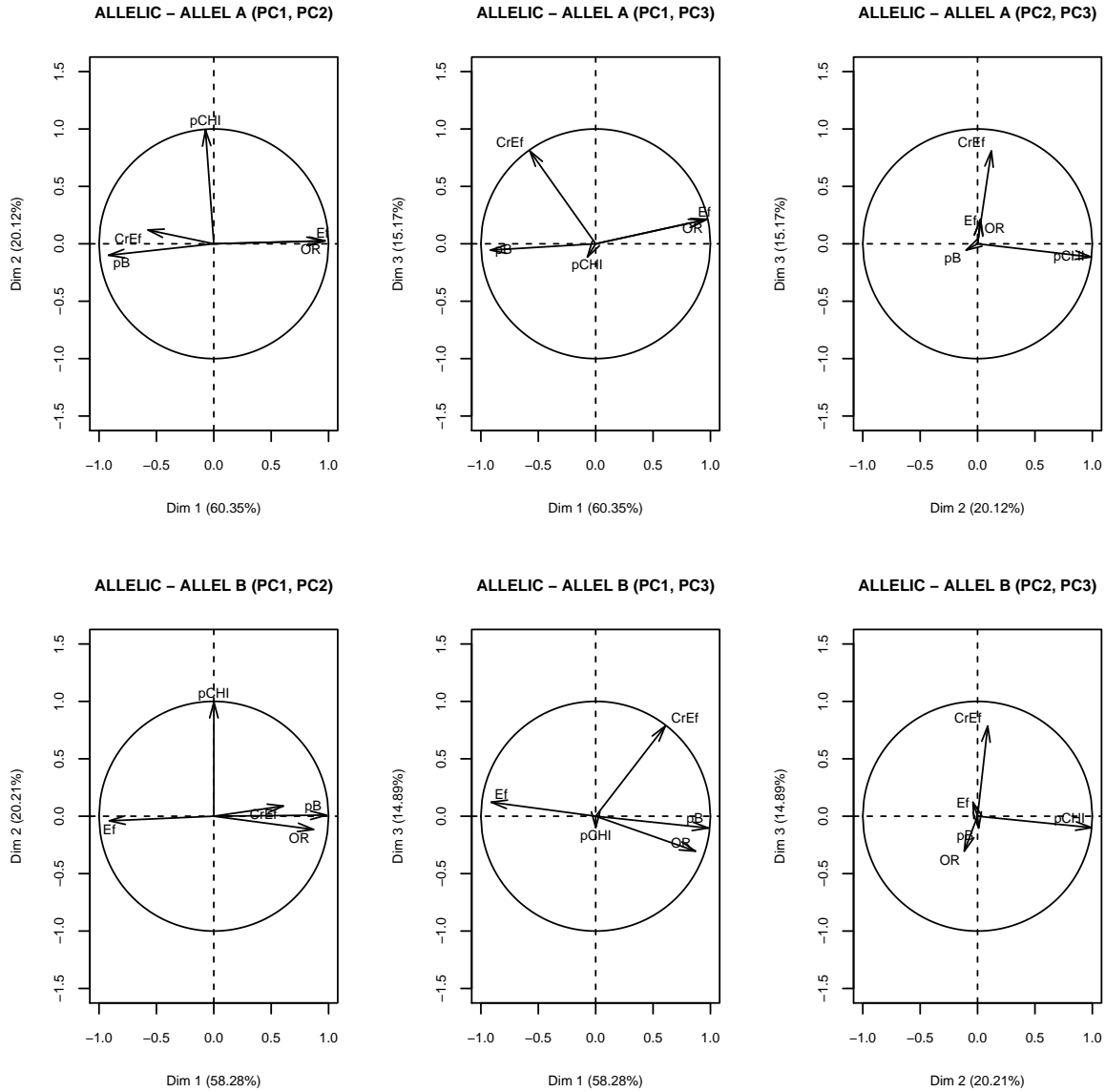


Figure 6: The weights of the parameters in each Principal Component (the first with the second, the first with the third and the second with the third, respectively), taken for the variables of the Allelic model of CT is compared to the Allele A and B models of BT. It describes the eigenvalues (in %) of each principal component.

Table VII: Top smallest p -values computed with the BT for the Alleles A and B models ($< 10^{-5}$) in the TGEN dataset, described above (* symbol represents that several markers in LD have been omitted in a single signal with the lowest p -value).¹

CHR	BP	SNP	MAF	A1	A2	p _v BT	BT-Allele	φ_A	φ_B	$Psi_A^{0.05}$	$Psi_B^{0.05}$	p _v CT	OR_a	M_A	M_B
19	50114786	rs4420638	0,308	G	A	1,33E-42	1	2,44	0,72	2,12	0,76	2,21E-34	3,39598	754	1694
3	52481466*	rs6784615	0,069	C	T	9,43E-08	1	2,30	0,95	1,64	0,97	2,09E-06	2,43106	170	2278
5	121942614	rs11953981	0,058	G	A	9,60E-07	1	2,45	0,95	1,70	1,03	1,83E-05	2,56819	128	2086
12	93847903	rs249152*	0,193	A	G	2,64E-06	1	1,49	0,91	1,26	0,96	1,16E-05	1,63099	464	1938
2	51804001	rs17864593*	0,016	A	G	3,02E-06	1	5,40	0,98	2,24	1,00	NA	5,51085	39	2329
2	205265992	rs41511746	0,017	G	C	3,76E-06	2	0,25	1,02	0,47	1,01	7,82E-06	0,241278	42	2408
16	26556972	rs12162084	0,157	A	G	4,10E-06	2	0,66	1,08	0,80	1,04	9,34E-06	0,611134	385	2063
1	21773864	rs1536934	0,069	A	G	4,21E-06	2	0,51	1,05	0,69	1,02	9,17E-06	0,479325	153	2057
10	68271216	rs4486514	0,04	C	T	4,30E-06	1	2,71	0,97	1,69	1,01	7,54E-05	2,80962	93	2231
10	112534645	rs7077757*	0,211	T	C	5,22E-06	2	0,71	1,10	0,83	1,05	1,21E-05	0,645986	515	1925
15	90462008	rs11074041*	0,142	C	G	5,45E-06	2	0,64	1,08	0,78	1,03	1,23E-05	0,597382	335	2019
18	71895842	rs359739*	0,192	A	C	6,64E-06	1	1,46	0,92	1,22	0,95	2,58E-05	1,59007	470	1976
8	47534249	rs4313171	0,067	T	C	7,68E-06	2	0,53	1,05	0,71	1,03	1,65E-05	0,502977	165	2285
22	17002691	rs12168275	0,038	G	C	8,38E-06	1	2,72	0,97	1,65	1,00	0,00013	2,80573	87	2215
5	117744488	rs6595122	0,282	C	A	8,61E-06	2	0,75	1,12	0,86	1,06	2,06E-05	0,667943	637	1621
19	50323656	rs17643262*	0,082	A	G	9,07E-06	1	1,88	0,95	1,37	0,97	5,69E-05	1,97991	192	2148
9	5583190	rs10815248	0,042	A	T	9,23E-06	1	2,49	0,97	1,58	0,99	0,00011	2,57753	101	2309
10	53698470	rs10824310	0,065	T	C	9,42E-06	2	0,52	1,05	0,71	1,03	2,02E-05	0,500702	159	2291
16	77974064	rs7192960	0,129	T	C	9,57E-06	2	0,64	1,07	0,79	1,04	2,12E-05	0,59958	316	2134

¹ The columns represent from left to right the number of chromosome (Chr), the base pair position of the SNP (BP), the name of the SNP (SNP), the MAF value, code for minor allele (A1), code for the other allele (A2), the lowest p -value of BT allelic models (p_vBT), the number of the allele model with lowest p -value of BT allelic models (BT-Allele), the effect of BT allelic models (φ_A and φ_B), the Critical effect of BT allelic models ($Psi_A^{0.05}$ and $Psi_B^{0.05}$), the p -value of the CT (p_vCT), the odds-ratio of this model (OR_a) and the number of individuals which present this alleles (M_A and M_B).

located 14 kilobase pairs distal to the APOE epsilon variant as the major finding. This observation was previously reported by TGEN researchers (Coon et al. [2007, Apr]). APOE locus is the most important genetic risk factor for Alzheimer's disease reported to date (Corder et al. [1993, Aug]). Notably, we found APOE locus significance more than eight orders of magnitude smaller using BT compared to CT (Table VII).

The rest of top markers also display smaller p -values using BT compared to CT calculations. This can be explained by the fact that the CT is a one-tailed test where the null hypothesis of no association is rejected if p -value turns out to be less than ϵ , while the BT is a two-tailed test (where the risk or protective role of the SNP is known by the p -value) and the rejection area for BT is one half that for CT ($\epsilon/2$).

Notice that the non-available p -value for the CT in the Chromosome 2 is due to the lack of data in a given cell. Anyway, this can be corrected by the Fisher test, which is not generally recommended due its computational cost.

5 Discussion

Any description of allelic distributions in the genome must begin by constructing a model of allelic probabilities. However, this important point remains unaddressed in many scientific literature, at least to the best of our knowledge. Indeed, almost all simulations, made for testing

a GWAS method, assume that allelic frequencies follow a uniform distribution. Here a model for allelic probability distribution is proposed and tested. In addition, we also improved the commonly used uniform distribution model.

The proposed alleles probability model, \mathcal{LAP}_t , offers a common scenario for each data set, characterizing noise. The truncation when not known can be estimated using the empirical distribution of the allelic probabilities. Regardless of the truncation, the remaining noise (quality control, stratification, insufficient population, etc...) is gathered with the variable \mathcal{D}_t .

The model depends only on the population of cases, controls, and the number of occurrences of the feature in the sample. Although the study was focused in a univariate analysis and s can be taken as the genotype in a single SNP, note that s is, in general, a vector and can represent any desired condition. For instance, s can include, along the genotype, the sex, age or other information. Furthermore, this vector can also include more than one genotype whether considering their interactions (epistasis) or not.

Although the examples used in this work described the allelic model, any other models could also be also analyzed without further modification of the method.

The new genome-wide association method (BT) has some commonalities with the conventional one. However, BT offers a remarkable ranking variability that might represent genuine signals. Novel candidate must be corroborated by intensive replication, multiple testing control, and meta-analysis using other datasets. We are aware that if BT can isolate novel loci, which generally missed while using traditional approaches, its application may help to uncover a fraction of the missing piece of heritability still pending for multiple complex traits. Consequently, next step of our research would be the generation of well powered meta-analysis of BT rankings. The isolation of genome-wide significant signals, and ultimately, the replication in independent series may help to measure the utility of this novel GWAS approach.

A Allelic probabilities variations

The difference between truncated universal allelic probability, \mathcal{AP}_t , and a priori truncated local allelic probability, \mathcal{LAP}_t , can be described in a single expression, $\mathcal{D}_t = \mathcal{AP}_t - \mathcal{LAP}_t$ the divergence of local MAF from commercial chips truncation for small MAF. Since the expected values of \mathcal{AP}_t and \mathcal{LAP}_t should be equal, we assumed that the expected value of \mathcal{D}_t as zero, and its seemed a plausible assumption that \mathcal{D}_t follows a normal distribution. However, this would give rise to local allelic probabilities out of the interval $[0, 1]$. To avoid this, we assumed that the values of \mathcal{AP}_t in the SNPs present in the chip belong to the interval $[t, 1 - t]$, and the observed allelic probability takes values in $[0, 1]$. Therefore, the observed values of \mathcal{D}_t belong to the interval $[0, 1]$. Hence, we modeled \mathcal{D}_t as a truncated normal distribution $NT(0, \delta, -t, t)$. Its probability density function is given by

$$h(x) = \begin{cases} \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\delta^2}}}{\int_{-t}^t \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\delta^2}}}, & \text{if } -t < x < t; \\ 0, & \text{otherwise,} \end{cases} \quad (17)$$

The mean of \mathcal{D}_t is 0 and its variance is given by

$$\sigma^2(\mathcal{D}_t) = \sigma_D^2 = \delta^2 \left(1 - \frac{2\frac{t}{\delta}(\frac{t}{\delta})}{2\Phi(\frac{t}{\delta}) - 1} \right)$$

where ϕ and Φ are the PDF and CDF of the standard normal distribution $N(0, 1)$, respectively.

If we take $x = \frac{t}{\delta}$ then

$$\sigma_D^2 = t^2 \left(\frac{1 - \frac{2x\phi(x)}{2\Phi(x)-1}}{x^2} \right) = t^2 \left(\frac{1}{x^2} - \frac{2\phi(x)}{x(2\Phi(x)-1)} \right).$$

As \mathcal{AP}_t and \mathcal{D}_t were assumed to be independent the density function of

$$\mathcal{LAP}_t = \mathcal{AP}_t + \mathcal{D}_t$$

is the convolution of the probability density functions of \mathcal{AP}_t and \mathcal{D}_t :

$$(g * h)(z) = \int_{-\infty}^{+\infty} g(z-x)h(x)dx \quad (18)$$

B Properties of truncated allelic distribution

As \mathcal{AP}_t is distributed as $\text{Beta}(\alpha, \alpha, t, 1-t)$, the mean of \mathcal{AP}_t should be $\mu_{\alpha,t} = 0.5$. This can be easily deduced from its PDF $g(a)$ (see (3)).

In order to calculate the variance of \mathcal{AP}_t , it is convenient to use the incomplete beta function, which is defined as

$$B(t; \alpha, \beta) = \int_0^t a^{\alpha-1}(1-a)^{\beta-1}da.$$

This function is related with the beta function $\text{Beta}(\alpha, \beta)$ using the regularized incomplete beta function:

$$I_t(\alpha, \beta) = \frac{B(t; \alpha, \beta)}{B(\alpha, \beta)}$$

They satisfy the following properties (Paris [2010])

$$\begin{aligned} B(\alpha+1, \beta) &= \frac{\alpha}{\alpha+\beta} B(\alpha, \beta) \\ I_t(\alpha, \beta) &= 1 - I_{1-t}(\beta, \alpha) \\ I_t(\alpha+1, \beta) &= I_t(\alpha, \beta) - \frac{t^\alpha(1-t)^\beta}{\alpha B(\alpha, \beta)} \end{aligned} \quad (19)$$

By (1), the variance σ_α^2 of $\text{Beta}(\alpha, \alpha)$ is $\frac{1}{4(2\alpha+1)}$. Using (19) we calculated the variance of the truncated beta distribution:

$$\begin{aligned} \sigma_{\alpha,t}^2 &= \int_t^{1-t} \frac{a^{\alpha+1}(1-a)^{\alpha-1}}{\int_t^{1-t} r^{\alpha-1}(1-r)^{\alpha-1}dr} dr - \mu_{\alpha,t}^2 \\ &= \frac{B(\alpha+2, \alpha)(I_{1-t}(\alpha+2, \alpha) - I_t(\alpha+2, \alpha))}{B(\alpha, \alpha)(I_{1-t}(\alpha, \alpha) - I_t(\alpha, \alpha))} - \frac{1}{4} \\ &= \frac{\alpha+1}{4\alpha+2} \frac{\frac{t^\alpha(1-t)^\alpha(4t-2)}{(\alpha+1)B(\alpha, \alpha)} + 1 - 2I_t(\alpha, \alpha)}{1 - 2I_t(\alpha, \alpha)} - \frac{1}{4} \\ &= \frac{1}{(4\alpha+2)B(\alpha, \alpha)} \frac{t^\alpha(1-t)^\alpha(4t-2)}{1 - 2I_t(\alpha, \alpha)} + \frac{\alpha+1}{4\alpha+2} - \frac{1}{4} \end{aligned}$$

Therefore, the variance of \mathcal{AP}_t is

$$\sigma_u^2 = \sigma_{\alpha,t}^2 = \frac{1}{(4\alpha+2)\text{B}(\alpha,\alpha)} \frac{t^\alpha(1-t)^\alpha(4t-2)}{1-2I_t(\alpha,\alpha)} + \frac{\alpha+1}{4\alpha+2} - \frac{1}{4}. \quad (20)$$

Next, we needed the following technical lemma.

Lemma B.1. *Let $f(x) = \frac{1}{x^2} - \frac{2\phi(x)}{x(2\Phi(x)-1)}$. Then $f(x)$ is decreasing for $x > 0$, $f(x) < \frac{1}{3}$ for every $x > 0$ and $\lim_{x \rightarrow 0} f(x) = \frac{1}{3}$.*

Proof. First, we need to prove that $f(x) < 0$, if $x > 0$. For that, we consider the function $g(x) = 2\Phi(x) - 1 + \frac{6x\phi(x)}{x^2-3}$. Having in mind that $\phi'(x) = -x\phi(x)$ and $\Phi'(x) = \phi(x)$, we have

$$g'(x) = 2\phi(x) \left(1 + 3 \frac{(1-x^2)(x^2-3) - 2x^2}{(x^2-3)^2} \right) = -\frac{4x^4\phi(x)}{(x^2-3)^2} \leq 0.$$

Thus, g is strictly decreasing function in each region where it is continuous, for example, the interval $(0, \sqrt{3})$. Let $0 < x < \sqrt{3}$. As $g(0) = 0$, we deduce that $g(x) < 0$. As $x^2 - 3 < 0$, we have $0 < (x^2 - 3)g(x) = (2\Phi(x) - 1)(x^2 - 3) + 6x\phi(x)$. Plainly $(2\Phi(x) - 1)(x^2 - 3) + 6x\phi(x)$ is also positive if $x \geq \sqrt{3}$. This proves that $(2\Phi(x) - 1)(x^2 - 3) + 6x\phi(x) > 0$ for every $x > 0$. Using that $2\Phi(x) - 1, x^2 > 0$, we deduce that $\frac{3-x^2}{3x^2} - \frac{2\phi(x)}{x(2\Phi(x)-1)} < 0$ and hence

$$f(x) = \frac{1}{x^2} - \frac{2\phi(x)}{x(2\Phi(x)-1)} = \frac{1}{3} + \frac{3-x^2}{3x^2} - \frac{2\phi(x)}{x(2\Phi(x)-1)} < \frac{1}{3},$$

as desired.

Using L'Hôpital rule, we have $\lim_{x \rightarrow 0} \frac{2\Phi(x)-1}{x\phi(x)} = \lim_{x \rightarrow 0} \frac{2}{1-x^2} = 2$. Applying L'Hôpital again, we have

$$\begin{aligned} \lim_{x \rightarrow 0} \left(\frac{1}{x^2} - \frac{2\phi(x)}{x(2\Phi(x)-1)} \right) &= \lim_{x \rightarrow 0} \frac{2\Phi(x) - 1 - 2x\phi(x)}{x^2(2\Phi(x)-1)} \\ &= \lim_{x \rightarrow 0} \frac{x\phi(x)}{2\Phi(x) - 1 + x\phi(x)} = \lim_{x \rightarrow 0} \frac{1}{\frac{2\Phi(x)-1}{x\phi(x)} + 1} = \frac{1}{3} \end{aligned}$$

To prove that $f(x)$ is decreasing for $x > 0$, we consider the following functions

$$\begin{aligned} \psi(x) &= 2\Phi(x) - 1 \\ \alpha(x) &= (x^4 - 2x^2 + 3)\sqrt{x^4 + 2x^2 + 9} + x^6 - x^4 + 5x^2 - 9. \\ h(x) &= 2\psi(x) - x\phi(x)(x^2 + 1 + \sqrt{x^4 + 2x^2 + 9}). \end{aligned}$$

We claim that $\alpha(x) > 0$ for every $x > 0$. For that we use the following equality

$$(x^4 - 2x^2 + 3)^2(x^4 + 2x^2 + 9) - (x^6 - x^4 + 5x^2 - 9)^2 = 32x^4. \quad (21)$$

Therefore

$$(x^4 - 2x^2 + 3)^2(x^4 + 2x^2 + 9) > (x^6 - x^4 + 5x^2 - 9)^2.$$

Moreover, $x^4 - 2x^2 + 3 = (x^2 - 1)^2 + 2 > 0$ and, if $0 < x < 1$ then $x^6 - x^4 + 5x^2 - 9 < 0$. Thus, if $0 < x < 1$ then $(x^4 - 2x^2 + 3)\sqrt{x^4 + 2x^2 + 9} > -x^6 + x^4 - 5x^2 + 9$, or equivalently $\alpha(x) > 0$. This equality also holds for $x > 1$ because both $x^4 - 2x^2 + 3$, $x^4 + 2x^2 + 9$ and $x^6 - x^4 + 5x^2 - 9$

are increasing for $x > 1$. Therefore if $x > 1$ then $\alpha(x) \geq \alpha(1) = 2\sqrt{12} - 4 > 0$. This proves the claim.

Let $x > 0$. As

$$\begin{aligned} h'(x) &= \phi(x) \left(4 + x^2(x^2 + 1 + \sqrt{x^4 + 2x^2 + 9}) - \left(3x^2 + 1 + \sqrt{x^4 + 2x^2 + 9} + \frac{x(4x^3 + 4x)}{2\sqrt{x^4 + 2x^2 + 9}} \right) \right) \\ &= \frac{\phi(x)}{\sqrt{x^4 + 2x^2 + 9}} \left((x^4 - 2x^2 + 3)\sqrt{x^4 + 2x^2 + 9} + (x^2 - 1)(x^4 + 2x^2 + 9) - 2x^2(x^2 + 1) \right) \\ &= \frac{\phi(x)}{\sqrt{x^4 + 2x^2 + 9}} \left((x^4 - 2x^2 + 3)\sqrt{x^4 + 2x^2 + 9} + x^6 - x^4 + 5x^2 - 9 \right) \\ &= \frac{\phi(x)\alpha(x)}{\sqrt{x^4 + 2x^2 + 9}} \end{aligned}$$

we conclude that h is an increasing function, therefore, $h(x) > h(0) = 0$. Equivalently

$$\psi(x) > \frac{x\phi(x)(x^2 + 1 + \sqrt{x^4 + 2x^2 + 9})}{2}. \quad (22)$$

As the greatest root of the quadratic polynomial $q(T) = T^2 - (x^3 + x)\phi(x)T - 2x^2\phi(x)^2$ is $\frac{x\phi(x)(x^2 + 1 + \sqrt{x^4 + 2x^2 + 9})}{2}$, inequality (22) implies that $q(\psi(x)) > 0$. Hence

$$f'(x) = -\frac{2}{x^3} - 2\frac{-x^2\phi(x)\psi(x) - \phi(x)\psi(x) - 2x\phi(x)^2}{x^2\psi(x)^2} = -\frac{2q(\psi(x))}{x^3\psi(x)^2} < 0.$$

Thus f is decreasing for $x > 0$, as desired. \square

By Lemma B.1 we have $0 \leq \sigma_D^2 \leq \frac{t^2}{3}$.

As \mathcal{AP}_t and \mathcal{D}_t are independent, the variance of $\mathcal{LAP}_t = \mathcal{AP}_t + \mathcal{D}_t$ is $\sigma_l^2 = \sigma_u^2 + \sigma_D^2$, hence,

$$\sigma_u^2 \leq \sigma_l^2 \leq \sigma_u^2 + \frac{t^2}{3}.$$

Notice that the case $\sigma_l^2 = \sigma_u^2 + \frac{t^2}{3}$ occurs for the highest degree of noise, i.e., when \mathcal{D}_t has maximum variance $\sigma_D^2 = \frac{t^2}{3}$.

C The effects of a genotype s : $\varphi_s, \Psi_s^\epsilon$

We defined the discrete random variable \mathcal{F} , which takes only two values 0 and 1 depending on whether the individual is a control or a case, respectively. Let M_{0s} and M_{1s} be the number of controls and cases with the given genotype s , respectively, and let $M_s = M_{0s} + M_{1s}$. Let b_s denote the theoretical probability of being a case for an individual with genotype s in a GWAS with N_0 controls and N_1 cases.

We assumed that the presence of the genotype multiply by a factor φ_s as the probability of presenting a phenotype. We denoted the probability of having the phenotype for the individuals with genotype s in the general population as c_s . Let p be the expected value of c_s , which is usually called the prevalence of the phenotype in the general population. Therefore,

$$p\varphi_s = c_s = P(\mathcal{F} = 1|s) = \frac{P(\mathcal{F} = 1, s)}{P(s)} = \frac{P(\mathcal{F} = 1)P(s|\mathcal{F} = 1)}{P(s)} = \frac{pP(s|\mathcal{F} = 1)}{P(s)}$$

Since controls are a representation of the population and cases are a sample of individuals with the phenotype, then we can compute the following estimations:

$$b_s \approx \frac{M_{1s}}{M_s}, \quad P(s) \approx \frac{M_{0s}}{N_0} \quad \text{and} \quad P(s|\mathcal{F} = 1) \approx \frac{M_{1s}}{N_1}$$

Therefore

$$\varphi_s \approx \frac{N_0}{N_1} \frac{M_{1,s}}{M_{0,s}} = \frac{N_0}{N_1} \frac{M_{1,s}/M_s}{1 - M_{1,s}/M_s} = \frac{N_0}{N_1} \frac{\hat{b}_s}{1 - \hat{b}_s}$$

and so

$$\hat{b}_s = \frac{c_s N_1}{p N_0 + c_s N_1} = \frac{1}{\frac{p N_0}{c_s N_1} + 1} = \frac{\varphi_s N_1}{N_0 + \varphi_s N_1}.$$

Next, we proceeded as in the tests statistics defined in (10) for the new b_s that takes into account the effect of the genotype s . Therefore, we may construct similar statistics in the same way as above, where the decision rule at a $100(1 - \epsilon/2)\%$ are

$$\begin{aligned} &\text{Accept } H'_0 \quad \text{if} \quad \omega'_{\epsilon/2, \varphi} \leq \hat{b}_s \leq \omega'_{1-\epsilon/2, \varphi} \\ &\text{Reject } H'_0 \quad \text{otherwise} \end{aligned} \tag{23}$$

where $Pr(\text{Beta}(\alpha'_{M_s}, \beta'_{M_s}) < \omega'_{\epsilon/2, \varphi}) = Pr(\text{Beta}(\alpha'_{M_s}, \beta'_{M_s}) > \omega'_{1-\epsilon/2, \varphi}) = \epsilon/2$, and

$$\alpha'_{M_s} = \frac{\varphi N_1}{N_0 + \varphi N_1} \left(\frac{M_s(N_0 + \varphi N_1 - 1)}{N_0 + \varphi N_1 - M_s} - 1 \right) \quad \text{and} \quad \beta'_{M_s} = \frac{N_0}{N_0 + \varphi N_1} \left(\frac{M_s(N_0 + \varphi N_1 - 1)}{N_0 + \varphi N_1 - M_s} - 1 \right)$$

Here, we defined a new parameter, called *the critical effect* of the genotype s , with a certain confidence level ϵ as the effect value, φ_s , of the expected distribution under the decision rule (13) where $b_s = q'_{1-\epsilon/2}$ if $b_s > \hat{b} = \frac{N_1}{N}$ or $b_s = q'_{\epsilon/2}$ if $b_s \leq \hat{b} = \frac{N_1}{N}$.

That is, the critical effect size Ψ_s^ϵ of genotype s at a ϵ confidence level is the greatest real number φ_s (respectively lowest), such that the test given by (10) for $b_s = \frac{\varphi_s N_1}{N_0 + \varphi N_1}$ rejects the null hypothesis.

D Relationships among the models

Figure 7 shows how the models of CT and BT could be connected. Each considered variable (AA , BB or AB) was treated as a different model. This implies that there is no one-to-one correspondence between CT and BT models. For instance, the allelic CT model is not equal to the probability of being case with the allele A or B , as analyzed in BT model. On the contrary, both cases (A and B) must be taken into account when compared to the allelic CT model.

Thus, there are eight models in the BT which seems to be of interest: allele A (A), allele B (B), dominant of A ($AA \cup AB$), recessive of A (AA), dominant of B ($AB \cup BB$), recessive of B (BB), homozygous ($AA \cup BB$) and heterozygous (AB). Thus, summarizing that there are five interesting models for CT against eight meaningful models for the new BT.

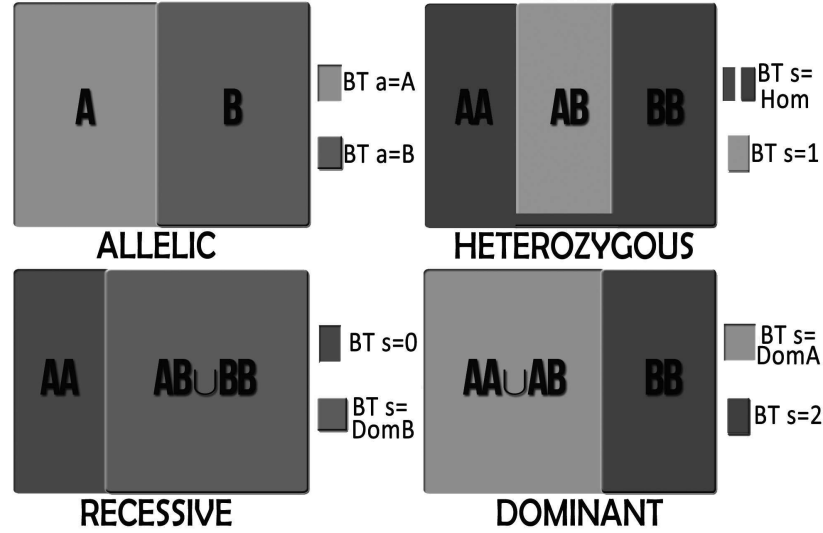


Figure 7: Paired models, *i.e.* relationships between the models of Conventional Test (CT), composed by allelic, heterozygous, recessive, and dominant; the new test proposed, called the Beta Test (BT), composed by allele A (BT $a = A$), allele B (BT $a = B$), $s = AA$ (BT $s = 0$), $s = AB$ (BT $s = 1$), $s = BB$ (BT $s = 2$), dominant of A (BT $s = DomA$), dominant of B (BT $s = DomB$), and homozygous (BT $s = Hom$).

References

- DM Altshuler, RA Gibbs, L Peltonen, and et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8, 2010.
- C Antúnez, M Boada, A González-Pérez, and et al. The membrane-spanning 4-domains, subfamily a (ms4a) gene cluster contains a common variant associated with alzheimer’s disease. *Genome Medicine.*, (3(5)), 2011, May 31.
- J Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1):37–46, 1960.
- KD Coon, AJ Myers, DW Craig, and et al. A high-density whole-genome association study reveals that apoe is the major susceptibility gene for sporadic late-onset alzheimer’s disease. *J Clin Psychiatry*, 68(4):613–8, 2007, Apr.
- EH Corder, AM Saunders, WJ Strittmatter, and et al. Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer’s disease in late onset families. *Science*, 13(261(5123)):921–3, 1993, Aug.
- PI de Bakker, MA Ferreira, X Jia, and et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet*, 17(R2):R122–8, 2008.
- KA Frazer, DG Ballinger, DR Cox, and et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–61, 2007.

- LA Hindorff, P Sethupathy, HA Junkins, and et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23):9362–7, 2009.
- L Isserlis. On the value of a mean as calculated from a sample. *Journal of the Royal Statistical Society*, 81 (1):75–81, 1918.
- ES Lander, LM Linton, B Birren, and et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.
- D MacKay. Information theory, inference and learning algorithms. *Cambridge University Press; First Edition.*, (ISBN 978-0521642989), 2003.
- TA Manolio, FS Collins, NJ Cox, and et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, 2009.
- C Meesters, M Leber, C Herold, and et al. Quick, "imputation-free" meta-analysis with proxy-snps. *BMC bioinformatics*, 13:231, 2012.
- R. B. Paris. *NIST Handbook of Mathematical Functions*, Cambridge University Press. *Incomplete beta functions*. Cambridge University Press, 2010.
- K Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, (2 (11)):559–572, 1901.
- EM Reiman, JA Webster, AJ Myers, and et al. Gab2 alleles modify alzheimer’s risk in apoe epsilon4 carriers. *Neuron*, (54):713–720, 2007.
- M Ruiz-Marín, M Matilla-García, JA García-Córdoba, and et al. An entropy test for single-locus genetic association analysis. *BMC Genetics*, (11-19), 2010.
- R Sachidanandam, D Weissman, SC Schmidt, and et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–33, 2001.
- R Shelton and JA Cliffe. Spherical cows. 2007. URL http://imagine.gsfc.nasa.gov/docs/features/topics/snr_group/spherical_cow.html.